INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Finite mixture models for exponential repeated data

Christian Lavergne — Marie-José Martinez — Catherine Trottier

N° 6119

Février 2007

Thème BIO

ISSN 0249-6399 ISRN INRIA/RR--6119--FR+ENG

Finite mixture models for exponential repeated data

Christian Lavergne, Marie-José Martinez*, Catherine Trottier

Thème BIO — Systèmes biologiques
Projet VirtualPlants

Rapport de recherche n° 6119 — Février 2007 — 18 pages

Abstract: The analysis of finite mixture models for exponential repeated data is considered. The mixture components correspond to different possible states of the statistical units. Dependency and variability of repeated data are taken into account through random effects. For each component, an exponential mixed model is thus defined. When considering parameter estimation in this mixture of exponential mixed models, the EM algorithm cannot be directly used since the marginal distribution of each mixture component cannot be analytically derived. In this paper, we propose two parameter estimation methods. The first one uses a linearisation specific to each exponential mixed model within each component. The second approach uses a Metropolis-Hastings algorithm as a building block of an MCEM algorithm.

Key-words: Generalized linear model, Random effect, Mixture model, EM algorithm, Metropolis-Hastings algorithm

* Corresponding author. Email : martinez@math.univ-montp2.fr

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 —Télécopie : +33 4 92 38 77 65

Modèles de mélange fini pour des données exponentielles répétées

Résumé : Nous nous intéressons à un modèle de mélange pour des données répétées de loi exponentielle. Les composants du mélange traduisent différents états possibles des individus. Pour chacun de ces composants, on modélise la dépendance et l'extra-variabilité dues à la répétition des données par l'introduction d'effets aléatoires. Dans ce modèle de mélange exponentiel mixte, la distribution marginale n'étant pas accessible, l'utilisation de l'algorithme EM n'est pas directement envisageable. Nous proposons alors une première méthode d'estimation des paramètres basée sur une linéarisation spécifique à la loi exponentielle. Nous proposons ensuite une méthode plus générale puisque s'appuyant sur une étape de Metropolis-Hastings pour construire un algorithme de type MCEM. Cet algorithme est applicable pour un mélange de modèles linéaires généralisés mixtes quelconques.

Mots-clés : Modèle linéaire généralisé, Effet aléatoire, Modèle de mélange, Algorithme EM, Algorithme de Metropolis-Hastings

Contents

1 Introduction

In the past decades, finite mixture models have been extensively developed in the literature. Surveys on these issues can be found in Titterington, Smith and Makov [17], McLachlan and Basford [13] and McLachlan and Peel [15]. In such finite mixture models, it is assumed that a sample of observations arises from a specified number of underlying groups or classes with unknown proportions and according to a specific form of distribution in each of them. A large number of distributions from the exponential family have been considered such as normal, Poisson, exponential (Hasselblad [6]). Wedel and DeSarbo [18] have proposed a mixture of generalized linear models which contains the previously proposed mixtures as special cases, as well as a host of other parametric specifications theretofore not dealt with in the literature. The use of these mixture models can be, in particular, a way to consider unexpected variance in GLM's or also a way to take into account underlying unobserved latent variable forming groups or classes. More recently, Celeux, Martin and Lavergne [2] have proposed a mixture of linear mixed models (LMM) in a microarray data analysis context. The introduction of random effects allowed them to take into account the variability of gene expression profiles from repeated microarray experiments. In our work, we consider the analysis of finite mixture for exponential repeated data. The mixture components correspond to different possible states of the statistical units. Dependency and variability of exponential repeated data are taken into account through exponential mixed models defined for each mixture component. In the field of the Health Sciences, applications may concern the modelling of lengths of repeated hospital stays for patients belonging to unknown clusters. Another example is the analysis of elimination times after repeated absorptions of a drug by patients not being controlled a priori.

Concerning parameter estimation in the proposed mixture of exponential mixed models, the use of the EM algorithm which allows to take into account the incomplete structure of the data is considered (McLachlan and Krishnan [14]). But the algorithm presented by Celeux et al. [2] for a mixture of linear mixed models cannot here be transposed because the marginal distribution of each mixture component cannot be analytically derived. Thus, we propose two parameter estimation methods. The first one uses a linearisation specific to the exponential distribution hypothesis associated with each mixture component. The second approach is adapted from the algorithm presented by McCulloch [11] for generalized linear mixed models (GLMM) [12] and uses a Metropolis-Hastings step (Hastings [7]) to allow construction of an MCEM algorithm. This algorithm can be adapted to a mixture of any generalized linear mixed models. The paper is organized as follows. After a description of the model hypotheses in section 2, we outline the EM algorithm presented by Celeux et al. for a mixture of linear mixed models in section 3. In section 4, we describe the two proposed parameter estimation methods. Finally in section 5, we study the behaviour of these approaches on simulations.

2 Mixture of exponential mixed models : model definition

Consider $y = (y_1', \ldots, y_I')'$ a vector of observations where y_i is associated with the ith statistical unit. Each y_i contains the n_i repetitions y_{ij}. Consider also different components $\mathcal{C}_k$, $k = 1, \ldots, K$, corresponding to different possible states of the statistical units. We assume that all repeated measures of a statistical unit belong to the same component and we define the indicator vectors $z_i = (z_{i1}, \ldots, z_{iK})$, $i = 1, \ldots, I$, with $z_{ik} = 1$ if unit $i \in \mathcal{C}_k$ and 0 otherwise.

To take into account dependency and variability of repeated data, we consider for each component an exponential mixed model with a random effect associated with each unit. This leads to a mixture of exponential mixed models and the density of Y_i may be written as follows:

$$f(y_i|\theta, p) = \sum_{k=1}^{K} p_k f_k(y_i|\theta_k)$$

where the p_k's are mixing weights with $0 < p_k < 1$ for $k = 1, \ldots, K$ and $\sum_{k=1}^{K} p_k = 1$, and $f_k(.|\theta_k)$ denotes the density function of the marginal distribution associated with the exponential mixed model with unknown parameters $\theta_k = (\beta_k, \sigma_k^2)$. Note that this marginal distribution cannot be analytically derived.

More precisely, given the mixture component $\mathcal{C}_k$ from which unit i arises and given the unobserved random effect ξ_i, Y_i is assumed to be exponentially distributed:

$$(Y_i|\xi_i, Z_{ik} = 1) \sim \mathcal{E}xp(\mu_{\xi,i}^k) \text{ with } \left\{ \begin{array}{l} \mu_{\xi,i}^k = \exp(X_i\beta_k + U_i\xi_i) \\ (\xi_i|Z_{ik} = 1) \sim \mathcal{N}(0, \sigma_k^2) \end{array} \right.$$

where

- $\forall i, i' \in \{1, \ldots, I\}^2$ $i \neq i'$, ξ_i and $\xi_{i'}$ are assumed to be independent,
- β_k is the $q \times 1$ fixed effect parameter vector associated with component $\mathcal{C}_k$,
- σ_k^2 is the variance of the random effect associated with component $\mathcal{C}_k$,
- $X_i = \left(\begin{array}{c} \underline{x_{i1}'} \\ \vdots \\ \underline{x_{in_i}'} \end{array} \right)$ and $U_i = (u_{i1}, \ldots, u_{in_i})'$ are the $n_i \times q$ and $n_i \times 1$ known design matrices.

Thus, we focus here on a mixture model-based approach to the clustering of exponential repeated data (McLachlan and Peel [15]).

3 EM algorithm for a mixture of linear mixed models

In this section, we outline the maximum likelihood estimation approach for a mixture of linear mixed models using the EM algorithm presented by Celeux et al. [2]. The EM methodology takes into account the incomplete structure of the data (Dempster, Laird and Rubin [4]). Missing data are here of two types : the indicator vectors z_i, $i = 1, \ldots, I$ of unit memberships to the mixture components and the random effects ξ_i, $i = 1, \ldots, I$.

Given the mixture component $\mathcal{C}_k$ from which unit i arises, Y_i is here assumed to be normally distributed:

$$(Y_i|Z_{ik} = 1) \quad = \quad X_i\beta_k \; + \; U_i\xi_i \; + \; \varepsilon_i$$

where

- $(\xi_i|Z_{ik} = 1) \sim \mathcal{N}(0, \sigma_k^2)$,
- ε_i is the $n_i \times 1$ error vector assumed to be normally distributed: $\varepsilon_i \sim \mathcal{N}(0, \tau^2 \, I_{n_i})$ with I_{n_i} the identity matrix of order n_i.
- $\forall i \in \{1, \ldots, I\}$, ε_i and ξ_i are assumed to be independent and $\forall i, i' \in \{1, \ldots, I\}^2 \; i \neq i'$, ξ_i and $\xi_{i'}$, respectively ε_i and $\varepsilon_{i'}$, are assumed to be independent,
- β_k is the $q \times 1$ fixed effect parameter vector associated with component $\mathcal{C}_k$,
- σ_k^2 is the random effect variance associated with component $\mathcal{C}_k$,
- $X_i = \begin{pmatrix} \underline{x'_{i1}} \\ \vdots \\ \underline{x'_{in_i}} \end{pmatrix}$ and $U_i = (u_{i1}, \ldots, u_{in_i})'$ are the $n_i \times q$ and $n_i \times 1$ design matrices.

Thus the distribution of Y_i is a mixture of linear mixed models defined by

$$f(y_i|\theta, p) = \sum_{k=1}^{K} p_k f_k(y_i|\theta_k)$$

where $p = (p_1, \ldots, p_K)$ are the mixing weights, $\theta = (\theta_1, \ldots, \theta_K)$ with $\theta_k = (\beta_k, \sigma_k^2, \tau^2)$ the linear mixed model parameters associated with component $\mathcal{C}_k$, and $f_k(y_i|\theta_k)$ denotes the density function of the distribution of Y_i i.e. a Gaussian distribution with mean $X_i\beta_k$ and variance matrix $\Gamma_{k,i} = \tau^2 I_{n_i} + \sigma_k^2 U_i U_i'$. In their paper, Celeux et al. [2] consider a mixture model where all parameters are dependent on component $\mathcal{C}_k$. We consider here a mixture model where the parameters β_k and σ_k^2 depend on component $\mathcal{C}_k$ whereas the residual variance τ^2 is the same for all mixture components.

The log-likelihood associated with the complete data (y, z, ξ) is given by

$$L(\theta, p|y, z, \xi) = \sum_{i=1}^{I} \sum_{k=1}^{K} z_{ik} \Big\{ \ln p_k + \ln f(y_i, \xi_i|z_{ik} = 1, \theta_k) \Big\}$$

where $\ln f(y_i, \xi_i | z_{ik} = 1, \theta_k)$ can be written as

$$\ln f(y_i|\xi_i, z_{ik} = 1, \theta_k) + \ln f(\xi_i | z_{ik} = 1, \theta_k)$$

with

- $$\begin{aligned} \ln f(y_i|\xi_i, z_{ik} = 1, \theta_k) &= -\tfrac{1}{2}\big(n_i \ln 2\pi + n_i \ln \tau^2 + \frac{\varepsilon_i'\varepsilon_i}{\tau^2}\big), \\ &= -\tfrac{1}{2}\Big\{ n_i \ln 2\pi + n_i \ln \tau^2 \\ &\qquad + \tfrac{(y_i - X_i\beta_k - U_i\xi_i)'(y_i - X_i\beta_k - U_i\xi_i)}{\tau^2} \Big\}. \end{aligned}$$
- $$\ln f(\xi_i | z_{ik} = 1, \theta_k) = -\tfrac{1}{2}\big(\ln 2\pi + \ln \sigma_k^2 + \frac{\xi_i^2}{\sigma_k^2}\big)$$

At iteration $[t+1]$, the E step consists of computing the expectation of the complete data log-likelihood given the observed data and a current value of the parameters ($\theta^{[t]}$, $p^{[t]}$):

$$\begin{aligned} Q(\theta, p|\theta^{[t]}, p^{[t]}) &= E\Big[L(\theta, p|y, z, \xi)|y, \theta^{[t]}, p^{[t]}\Big] \\ &= \sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i) \ln p_k - \frac{1}{2}\sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i)\Big\{(n_i+1)\ln 2\pi \\ &\qquad + n_i \ln \tau^2 + \ln \sigma_k^2 + \frac{E_{Ck}^{[t]}(\varepsilon_i'\varepsilon_i)}{\tau^2} + \frac{E_{Ck}^{[t]}(\xi_i^2)}{\sigma_k^2}\Big\} \end{aligned}$$

where $E_{Ck}^{[t]}(.) = E(.|y_i, z_{ik} = 1, \theta_k^{[t]})$

$$\begin{aligned} \text{and} \quad t_k^{[t]}(y_i) &= P(Z_{ik} = 1|y_i, \theta^{[t]}, p^{[t]}) \\ &= \frac{p_k^{[t]}\, f(y_i|z_{ik} = 1, \theta_k^{[t]})}{f(y_i|\theta^{[t]}, p^{[t]})} = \frac{p_k^{[t]}\, f_k(y_i|\theta_k^{[t]})}{\sum_{l=1}^{K} p_l^{[t]}\, f_l(y_i|\theta_l^{[t]})} \end{aligned}$$

denotes the posterior probability that unit i arises from component $\mathcal{C}_k$.

The M step consists of maximizing $Q(\theta, p|\theta^{[t]}, p^{[t]})$. It leads to the following explicit expressions for $k = 1, \ldots, K$:

$$p_k^{[t+1]} = \frac{\sum_{i=1}^{I} t_k^{[t]}(y_i)}{I}$$

$$\beta_k^{[t+1]} = \big(\sum_{i=1}^{I} t_k^{[t]}(y_i) X_i' X_i\big)^{-1} \sum_{i=1}^{I} t_k^{[t]}(y_i)\Big\{\tau^{2[t]}\, X_i'\, \Gamma_{k,i}^{[t]-1}\, (y_i - X_i\beta_k^{[t]}) + X_i'\, X_i \beta_k^{[t]}\Big\}$$

$$\sigma_k^{2[t+1]} = \frac{1}{\sum_{i=1}^{I} t_k^{[t]}(y_i)} \sum_{i=1}^{I} t_k^{[t]}(y_i)\Big\{\sigma_k^{4[t]}\,(y_i - X_i\beta_k^{[t]})'\,\Gamma_{k,i}^{[t]-1} U_i\,U_i'\,\Gamma_{k,i}^{[t]-1}\,(y_i - X_i\beta_k^{[t]}) + \sigma_k^{2[t]} - \sigma_k^{4[t]}\mathrm{tr}(\Gamma_{k,i}^{[t]-1} U_i U_i')\Big\}$$

$$\tau^{2[t+1]} = \frac{1}{n}\sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i)\Big\{\tau^{4[t]}(y_i - X_i\beta_k^{[t]})'\,\Gamma_{k,i}^{[t]-1}\,\Gamma_{k,i}^{[t]-1}\,(y_i - X_i\beta_k^{[t]}) + n_i\tau^{2[t]} - \tau^{4[t]}\mathrm{tr}(\Gamma_{k,i}^{[t]-1})\Big\}.$$

Details can be found in Celeux et al. [2].

4 Two parameter estimation methods

We consider here the parameter estimation for the mixture of exponential mixed models presented in section 2. In this context, the use of the EM algorithm is not directly possible. The complete data log-likelihood associated to this model is given by

$$\begin{aligned} L(\theta, p|y, \xi, z) &= \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\ln p_k + \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\ln f(y_i|\xi_i, z_{ik}=1, \theta_k) \\ &\quad + \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\ln f(\xi_i|z_{ik}=1,\theta_k) \end{aligned}$$

with

- $$\begin{aligned} \ln f(y_i|\xi_i, z_{ik}=1,\theta_k) &= \sum_{j=1}^{n_i} \ln f(y_{ij}|\xi_i, z_{ik}=1), \\ &= -\sum_{j=1}^{n_i}\Big\{x_{ij}'\beta_k + u_{ij}\xi_i + \frac{y_{ij}}{\exp(x_{ij}'\beta_k + u_{ij}\xi_i)}\Big\} \end{aligned}$$

because the y_{ij}'s are independent conditionally on ξ_i.

- $$\ln f(\xi_i|z_{ik}=1,\theta_k) = -\frac{1}{2}\big(\ln 2\pi + \ln\sigma_k^2 + \frac{\xi_i^2}{\sigma_k^2}\big)$$

In this case, at iteration $[t+1]$, the EM algorithm leads to formulae depending on conditional expectations $E_{Ck}^{[t]}(\xi_i^2)$, $E_{Ck}^{[t]}\big[\exp(-u_{ij}\xi_i)\big]$ and posterior probabilities $t_k^{[t]}(y_i)$, $i = 1,\ldots,I$, $k = 1,\ldots,K$. Because of the non-availability of the marginal distribution for each mixture component, probabilities $t_k^{[t]}(y_i)$ cannot be derived in closed form. Furthermore, neither the conditional expectation $E_{Ck}^{[t]}(\xi_i^2)$ nor $E_{Ck}^{[t]}\big[\exp(-u_{ij}\xi_i)\big]$ can be computed too since these calculations involve the unknown conditional distribution of ξ_i given y_i. We propose two parameter estimation methods which allow to get round these problems.

4.1 A method based on linearisation

This first approach is a conceptually simple method which involves two steps: a linearisation specific to the exponential mixed model (Gaudoin, Lavergne and Soler [5]) associated with each mixture component and the use of the EM algorithm for parameter estimation in a mixture of linear mixed models.

Knowing the component $\mathcal{C}_k$, the distribution associated with statistical unit i is given by

$$(Y_i|\xi_i, Z_{ik} = 1) \sim \mathcal{E}xp(\mu^k_{\xi,i}),$$

or equivalently :

$$\frac{Y_i}{\mu^k_{\xi,i}} \sim \mathcal{E}xp(1),$$

thus

$$\ln(Y_i) - \ln(\mu^k_{\xi,i}) \sim \mathcal{G}umbel,$$

where the Gumbel density function is defined by $\forall t \in I\!R\ f(t) = \exp(t - \exp(t))$ with mean $\gamma = -0.57722$ and variance $\dfrac{\pi^2}{6}$.
This enables us to write:

$$\ln(Y_i) - \ln(\mu^k_{\xi,i}) = \gamma + \varepsilon_i \quad \text{where} \quad E(\varepsilon_i) = 0_{n_i} \quad \text{and} \quad \text{var}(\varepsilon_i) = \frac{\pi^2}{6} I_{n_i}.$$

Defining the variable $D_i = \log(Y_i) - \gamma$, we end up with the linearized model:

$$D_i = X_i\beta_k + U_i\xi_i + \varepsilon_i$$

with 0-mean error vector ε_i and known variance matrix $\frac{\pi^2}{6} I_{n_i}$, which is viewed and considered as a linear mixed model $\mathcal{M}_k$ for the data $d_i = \log(y_i) - \gamma$ given the component $\mathcal{C}_k$.

Finally, we use the EM algorithm to estimate the parameters of the mixture of linear mixed models defined by

$$h(d_i|\theta, p) = \sum_{k=1}^{K} p_k h_k(d_i|\theta_k)$$

where $d_i = \ln(y_i) - \gamma$ and $h_k(d_i|\theta_k)$ is the density function of the Gaussian distribution with mean vector $X_i\beta_k$ and variance matrix $\Gamma_{k,i} = \frac{\pi^2}{6} I_{n_i} + \sigma^2_k U_i U_i'$. In this approach, note that vector d_i is derived from the data y_i whatever the component $\mathcal{C}_k$ from which unit i arises and without any use of the current value of the parameters.

The parameter estimation for this mixture of linear mixed models using the EM algorithm described in section 3 leads to the following expressions for $k = 1, \ldots, K$:

$$p_k^{[t+1]} = \frac{\sum_{i=1}^{I} t_k^{[t]}(d_i)}{I}$$

$$\beta_k^{[t+1]} = \Big(\sum_{i=1}^{I} t_k^{[t]}(d_i) X_i' X_i\Big)^{-1} \sum_{i=1}^{I} t_k^{[t]}(d_i) \Big\{ \frac{\pi^2}{6} X_i' \Gamma_{k,i}^{[t]-1} (d_i - X_i \beta_k^{[t]}) + X_i' X_i \beta_k^{[t]} \Big\}$$

$$\sigma_k^{2[t+1]} = \frac{1}{\sum_{i=1}^{I} t_k^{[t]}(d_i)} \sum_{i=1}^{I} t_k^{[t]}(d_i) \Big\{ \sigma_k^{4[t]} (d_i - X_i \beta_k^{[t]})' \Gamma_{k,i}^{[t]-1} U_i U_i' \Gamma_{k,i}^{[t]-1} (d_i - X_i \beta_k^{[t]}) + \sigma_k^{2[t]} - \sigma_k^{4[t]} \mathrm{tr}(\Gamma_{k,i}^{[t]-1} U_i U_i') \Big\}$$

4.2 An MCEM algorithm

The proposed algorithm is adapted from the MCEM algorithm presented by McCulloch [11] for generalized linear mixed models. Since expectations $E_{Ck}^{[t]}(\xi_i^2)$ and $E_{Ck}^{[t]}\big[\exp(-u_{ij}\xi_i)\big]$ and posterior probabilities $t_k^{[t]}(y_i)$ cannot be derived in closed form, our goal is to form Monte Carlo approximations of these quantities. To this aim, we incorporate a Metropolis-Hastings step into the EM algorithm which does not require specification of the marginal distribution of Y_i. This leads us to draw values from the unknown conditional distribution of ξ_i given Y_i, $Z_{ik} = 1$ and the current value $\theta_k^{[t]}$. One can then calculate Monte Carlo approximations of the two required expectations. In the same way, we draw values from the known distribution of ξ_i given $Z_{ik} = 1$ and the current value $\theta_k^{[t]}$ in order to approximate marginal distribution $f_k(y_i|\theta_k^{[t]})$ by Monte Carlo methods and to calculate posterior probability $t_k^{[t]}(y_i)$. Before presenting the proposed algorithm in section 4.2.2, we recall the Metropolis-Hastings step applied to our specific case in section 4.2.1.

4.2.1 The Metropolis-Hastings step

The Metropolis-Hastings algorithm [7] is certainly one of the most famous MCMC methods (Robert and Casella [16]). The aim of the MCMC methods is to generate samples from a target distribution π unavailable in closed form. To this end, a candidate distribution h (called the instrumental or proposal distribution) must be specified from which potential new values are drawn. Among samples generated from h, Metropolis-Hastings selects representative samples of the target distribution π using an acception/rejection method.

To define the proposed Metropolis-Hastings step, we need to specify the candidate distribution h. We propose to take h equal to the marginal distribution in class $\mathcal{C}_k$ of ξ_i given the current value $\theta_k^{[t]}$ (McCulloch [11]). Let $\xi_i^{[m]}$ be the previous draw from the conditional distribution of $\xi_i|Y_i, Z_{ik} = 1$ given the current value $\theta_k^{[t]}$. The probability of accepting the

new value ξ_i^* generated using the candidate distribution h is given by

$$\rho(\xi_i^{[m]}, \xi_i^*) = \min\Big\{1, \frac{f(\xi_i^*|y_i, z_{ik}=1, \theta_k^{[t]})h(\xi_i^{[m]})}{f(\xi_i^{[m]}|y_i, z_{ik}=1, \theta_k^{[t]})h(\xi_i^*)}\Big\}$$

where the second term simplifies to:

$$\begin{aligned}\frac{f(\xi_i^*|y_i, z_{ik}=1, \theta_k^{[t]})h(\xi_i^{[m]})}{f(\xi_i^{[m]}|y_i, z_{ik}=1, \theta_k^{[t]})h(\xi_i^*)} &= \frac{f(\xi_i^*|y_i, z_{ik}=1, \theta_k^{[t]})f(\xi_i^{[m]}|z_{ik}=1, \theta_k^{[t]})}{f(\xi_i^{[m]}|y_i, z_{ik}=1, \theta_k^{[t]})f(\xi_i^*|z_{ik}=1, \theta_k^{[t]})} \\ &= \frac{f(y_i|\xi_i^*, z_{ik}=1, \theta_k^{[t]})}{f(y_i|\xi_i^{[m]}, z_{ik}=1, \theta_k^{[t]})}.\end{aligned}$$

By choosing h equal to the random effects distribution, probability ρ is simplified since the obtained formula only involves the specification of the conditional distribution of Y_i given ξ_i and the component $\mathcal{C}_k$ from which unit i arises.

4.2.2 The proposed MCEM algorithm

Incorporating this Metropolis-Hastings step into the EM algorithm gives the following Monte Carlo EM (MCEM) algorithm at iteration $[t+1]$:

1. For $i = 1, \ldots, I$ and $k = 1, \ldots, K$, draw:

   - M values $\xi_i^{[1]}, \ldots, \xi_i^{[M]}$ from the distribution of $\xi_i|Y_i, Z_{ik} = 1$ given the current value $\theta_k^{[t]}$ using the Metropolis-Hastings algorithm described above and use them to form Monte Carlo approximations of the two required expectations in the function $Q(\theta, p|\theta^{[t]}, p^{[t]})$:

$$\begin{aligned}E_{Ck}^{[t]}(\xi_i^2) &\simeq \frac{1}{M}\sum_{m=1}^{M}\xi_i^{[m]2} \\ E_{Ck}^{[t]}\big[\exp(-u_{ij}\xi_i)\big] &\simeq \frac{1}{M}\sum_{m=1}^{M}\exp(-u_{ij}\xi_i^{[m]})\end{aligned}$$

   - N values $\xi_i^{[1]}, \ldots, \xi_i^{[N]}$ from the known distribution of ξ_i given $Z_{ik} = 1$ and the current value $\theta_k^{[t]}$ in order to approximate the marginal distribution:

$$\begin{aligned}f_k(y_i|\theta_k^{[t]}) &= f(y_i|z_{ik}=1, \theta_k^{[t]}) \\ &= \int\prod_{j=1}^{n_i} f(y_{ij}|\xi_i, z_{ik}=1, \theta_k^{[t]})\, f(\xi_i|z_{ik}=1, \theta_k^{[t]})\, d\xi_i \\ &\approx \frac{1}{N}\sum_{n=1}^{N}\Big\{\prod_{j=1}^{n_i} f(y_{ij}|\xi_i^{[n]}, z_{ik}=1, \theta_k^{[t]})\Big\}\end{aligned}$$

and to obtain an approximation of the posterior probability $t_k^{[t]}(y_i)$.

2. Then maximise the function $Q(\theta, p|\theta^{[t]}, p^{[t]})$ to obtain new parameter values $\theta^{[t+1]}$ and $p^{[t+1]}$.

5 Simulation results

5.1 Preliminary results

In order to study the behaviour of the MCEM algorithm developed in section 4.2, we first consider its use in the gaussian case. Indeed, in this case, the performances of the MCEM algorithm can easily be compared to those of the EM algorithm. We simulate 100 data sets. We give in Table 1 the mean and standard deviation of the 100 estimated values obtained with both EM and MCEM for each parameter. We consider here a two-component mixture model. We set the number of statistical units I equal to 100 and we consider the same number of repetitions for each unit: $\forall i = 1, \ldots, I \;\; n_i = J = 6$. The mixing parameters are $p_1 = 0.6$ and $p_2 = 0.4$. The random effect variances are $\sigma_1^2 = 0.2$ and $\sigma_2^2 = 0.8$ and the residual variance is $\tau^2 = 2$. We consider a unique fixed effect parameter by component: $\beta_1 = -2$ and $\beta_2 = 2$.

Table 1: Parameter estimation results obtained with EM and MCEM in the gaussian case on 100 simulated data sets

		EM		MCEM	
Simulated values		mean	s.d.	mean	s.d.
$\mathcal{C}_1$	$p_1 = 0.6$	0.6006	0.0171	0.6006	0.0170
	$\beta_1 = -2$	-1.9872	0.1141	-1.9873	0.1140
	$\sigma_1^2 = 0.2$	0.1994	0.1201	0.1991	0.1166
$\mathcal{C}_2$	$p_2 = 0.4$	0.3994	0.0171	0.3994	0.0170
	$\beta_2 = 2$	2.0289	0.1737	2.0292	0.1733
	$\sigma_2^2 = 0.8$	0.7657	0.2897	0.7605	0.2855
	$\tau^2 = 2$	2.0210	0.1340	2.0216	0.1339

Table 1 clearly shows that the MCEM algorithm performs close to the EM algorithm. However, it is important to note that the MCEM algorithm is numerically intensive. For instance, the EM algorithm implemented using R requires here only a few minutes whereas the MCEM algorithm implemented in C takes a few hours.

Table 2: Parameter estimation results from 100 simulated models defined by $\beta_1 = -3$ and $\beta_2 = 3$

		Model (A) $J = 4$		Model (A') $J = 8$	
Simulated values		Linear.	MCEM	Linear.	MCEM
	$p_1 = 0.6$	0.6017 (0.0039)	0.5999 (0.0023)	0.6002 (0.0012)	0.6001 (0.0010)
$\mathcal{C}_1$	$\beta_1 = -3$	-2.9990 (0.0878)	-2.9945 (0.0815)	-3.0154 (0.0921)	-3.0057 (0.0817)
	$\sigma_1^2 = 0.2$	0.2166 (0.1227)	0.1960 (0.0834)	0.1986 (0.0749)	0.1977 (0.0645)
	$p_2 = 0.4$	0.3983 (0.0039)	0.4001 (0.0023)	0.3998 (0.0012)	0.3999 (0.0010)
$\mathcal{C}_2$	$\beta_2 = 3$	3.0181 (0.1765)	3.0042 (0.1641)	3.0105 (0.1555)	3.0121 (0.1588)
	$\sigma_2^2 = 0.8$	0.7419 (0.2490)	0.7953 (0.2588)	0.7798 (0.2455)	0.7893 (0.2280)

5.2 Comparison of the two proposed methods

Simulations are performed to assess the ability of the proposed methods to estimate mixture parameters and to highlight the interest of taking into account repetitions.
We consider a two-component mixture model. We set the number of statistical units I equal to 100. The mixing parameters are $p_1 = 0.6$ and $p_2 = 0.4$ and the random effect variances are $\sigma_1^2 = 0.2$ and $\sigma_2^2 = 0.8$. We also consider one fixed effect parameter by component and we generate samples from:

- model (A) defined by $\beta_1 = -3$ and $\beta_2 = 3$,
- model (B) defined by $\beta_1 = -1$ and $\beta_2 = 1$.

It allows us to assess the ability of the methods to correctly estimate parameters when the mixture components are more (model (A)) or less separated (model (B)). In order to study the impact of the number of repetitions on the quality of the estimations, we also consider different number of repetitions J: we take $J = 4$ and $J = 8$.
Table 2 provides the mean and standard deviation of the estimations obtained from 100 samples generated from model (A). Table 3 gives the results obtained for model (B).

Table 3: Parameter estimation results from 100 simulated models defined by $\beta_1 = -1$ and $\beta_2 = 1$

		Model (B) $J = 4$		Model (B') $J = 8$	
Simulated values		Linear.	MCEM	Linear.	MCEM
	$p_1 = 0.6$	0.6634 (0.1340)	0.6536 (0.0851)	0.6481 (0.0856)	0.6330 (0.0797)
$\mathcal{C}_1$	$\beta_1 = -1$	-0.8908 (0.1845)	-0.9334 (0.1494)	-0.9485 (0.1599)	-0.9601 (0.1371)
	$\sigma_1^2 = 0.2$	0.2760 (0.2251)	0.2376 (0.1466)	0.2523 (0.1324)	0.2330 (0.1119)
	$p_2 = 0.4$	0.3366 (0.1340)	0.3464 (0.0851)	0.3519 (0.0856)	0.3670 (0.0797)
$\mathcal{C}_2$	$\beta_2 = 1$	1.2909 (0.5317)	1.2202 (0.3047)	1.2036 (0.3426)	1.1504 (0.2895)
	$\sigma_2^2 = 0.8$	0.5437 (0.4568)	0.6173 (0.3591)	0.6195 (0.3257)	0.6795 (0.3274)

Table 4: Correct classification rates (%) from 100 simulations

		Model (A) $\beta = (-3, 3)$ $J = 4$	Model (A') $\beta = (-3, 3)$ $J = 8$	Model (B) $\beta = (-1, 1)$ $J = 4$	Model (B') $\beta = (-1, 1)$ $J = 8$
Linear.	$\mathcal{C}_1$	99.98	100.00	93.37	96.38
	$\mathcal{C}_2$	99.55	99.95	67.87	76.95
MCEM	$\mathcal{C}_1$	99.96	100.00	96.15	96.52
	$\mathcal{C}_2$	99.92	99.97	73.17	79.95

Table 2 shows, in each situation ($J = 4$ and $J = 8$), that both methods provide accurate parameter estimations. As expected, the precision of the estimation depends on the random effect variances : the greater the variance, the greater the estimation's standard deviation. Note that the results obtained with the MCEM algorithm are slightly better than those obtained with the method based on linearisation. Table 2 also shows that the number of repetitions has an influence on the quality of the estimations.
The results shown in Table 3 are obtained when the mixture components are less separated. They are not as adequate as the first case but they are still reasonable. Remarks similar to those made for the first case can be made. We just note that the impact of the number of repetitions is more important in this case.

Table 4 provides the correct classification rate for each model using the maximum a posteriori (MAP) decision rule from the estimate parameter values $\hat{p}$, $\hat{\theta}$. The MAP decision rule consists of assigning all the measures of unit i to the mixture component $\mathcal{C}_k$ such as

$$k = \text{argmax}_{1 \leq l \leq K} \widehat{t_l(y_i)}$$

with $\widehat{t_l(y_i)} = P(Z_{il} = 1 | y_i, \hat{p}, \hat{\theta})$. The obtained results are globally satisfactory. Table 4 shows that the correct classification rate decreases when the random effect variance increases. It also clearly shows that the correct classification rate increases with the number of repetitions. Finally, we also note that the rates obtained with the MCEM algorithm are slightly better than those obtained with the method based on linearisation.

6 Discussion and conclusions

In this paper, we define a new class of models for repeated data: mixtures of generalized linear mixed models. These models allow us to introduce a notion of heterogeneity in the GLMM. They take dependency and variability of repeated data into account through random effects defined for each mixture component. We proposed two parameter estimation methods for these models: the MCEM algorithm which can be used for mixtures of any

generalized linear mixed models and the method based on linearisation specific to a mixture of exponential mixed models. These two methods are adaptations of the EM algorithm getting round problems related to the direct use of it.

The simulations performed in the exponential case show that the two proposed parameter estimation methods globally perform well. They also show that the MCEM algorithm gives slightly better results than the method based on linearisation. This behaviour difference is even greater in difficult situations with less separated mixture components or low number of repetitions. However, it is important to note that, like all MCMC approaches, the MCEM algorithm is numerically intensive since a large number of simulations is required at each iteration. In practice, a compiled programming language had to be used to reduce computation times. Moreover, even if this algorithm seems to perform well in practice, it is necessary to note that we still have not established theoretical results for convergence. On the contrary, the implementation of the method based on linearisation is fast and can easily be done with R for instance. Nevertheless, the use of this method is restricted to the exponential case.

Coming back to the numerically intensive problem of the MCEM algorithm, it would be interesting to propose an intermediate version using simulation via a stochastic approximation in order to avoid calculations. A future work could adapt the method developed by Kuhn and Lavielle [8]. In their paper, Kuhn and Lavielle proposed an algorithm combining the stochastic approximation version of EM (SAEM) [1] [3] with a Markov chain Monte Carlo procedure.

Finally, in this paper, it is assumed that the number of components is known. However, in practical situations, this is mostly not the case. It thus becomes a part of the estimation process. To determine the appropriate number of components, it would be interesting to consider the model selection problem for mixtures of generalized linear mixed models. Model selection criteria proposed by Martinez [10] and Lavergne et al. [9] for generalized linear mixed models could be adapted to mixtures of generalized linear mixed models.

References

[1] G. Celeux and J. Diebolt. A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and stochastics reports*, 41:119–134, 1992.

[2] G. Celeux, O. Martin, and C. Lavergne. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling*, 5:243–267, 2005.

[3] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, 27(1):94–128, 1999.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[5] O. C. Gaudoin, C. Lavergne, and J. L. Soler. A generalized geometric de-eutrophication software reliability model. *IEEE Transactions on Reliability*, 43:536–541, 1994.

[6] V. Hasselblad. Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association*, 64:1459–1471, 1969.

[7] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[8] E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM : Probability and Statistics*, 8:115–131, 2004.

[9] C. Lavergne, M.J. Martinez, and C. Trottier. Empirical model selection in generalized linear mixed effects models. *Computational Statistics (to appear)*, 2007.

[10] M.J. Martinez. *Modèles linéaires généralisés à effets aléatoires : contributions au choix de modèle et au modèle de mélange*. PhD thesis, Université Montpellier II, 2006.

[11] C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92:162–170, 1997.

[12] C. E. McCulloch and S. R. Searle. *Generalized, linear and mixed models*. Wiley Series in Probability and Statistics, 2001.

[13] G.J. McLachlan and K.E. Basford. *Mixture models: inference and application to clustering*. Marcel Dekker, New York, 1988.

[14] G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley, New York, 1997.

[15] G.J. McLachlan and D. Peel. *Finite mixture models*. Wiley-Interscience, New York, 2000.

[16] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer-Verlag, New York, second edition, 2004.

[17] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*. John Wiley & Sons Ltd, Chichester, 1985.

[18] M. Wedel and W. S. DeSarbo. A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12:21–55, 1995.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399